

Consistent Bandwidth Selection for Kernel Binary Regression

Naomi Altman *

Brenda MacGibbon †

Biometrics Unit

Département de mathématiques

Cornell University

Université du Québec à Montréal

September 19, 1995

BU-1126-MA

September 19, 1995

Abstract

The use of nonparametric regression techniques for binary regression is a promising alternative to parametric methods. As in other nonparametric smoothing problems, the choice of smoothing parameter is critical to the performance of the estimator and the appearance of the resulting estimate. In this paper, we discuss the use of selection criteria based on estimates of squared prediction risk and show asymptotic consistency and normality of the selected bandwidths. The usefulness of the methods are explored on a data set and in a small simulation study.

Keywords and phrases: Cross-validation; C_L ; U -statistics.

1991 AMS subject classification primary 62G07 secondary 62G20

*Supported by Hatch Grant 151410 NYF

†Supported by NSERC (Canada) and FCAR (Québec)

1 Introduction

Nonparametric regression can be particularly useful in the binary regression problem

$$y_i \sim \text{Bernoulli}[\mu(t_i)] \quad i = 1 \dots n \quad (1.1)$$

where $\mu(t)$ is a smooth function with square integrable p^{th} derivative, and t_i are fixed design points on $[0, 1]$ depending on the sample size n , since scatterplots of the raw data and of regression residuals are often difficult to interpret for binary response. Several authors have noted the advantages of smoothing for diagnostics in logistic regression (Altman, 1992; Azzalini, Bowman and Härdle, 1989; Copas, 1983; Fienberg and Gong, 1984; Fowlkes, 1987). However, a selection criterion for the choice of bandwidth for this problem has not been previously studied, although one has been suggested (Azzalini, Bowman and Härdle, 1989; Hastie and Tibshirani, 1990, p.159). In this article we examine the use of cross-validation (CV) (Allen, 1974; Geisser, 1975; Stone, 1974) and Mallows' C_L (Mallows, 1973) for selecting bandwidths for the nonparametric binary regression problem, using kernel regression estimators. Consistency of the estimators and convergence of the selected bandwidth is shown under average squared error loss.

Nonparametric regression techniques are frequently used for estimating a smooth mean function in the ordinary nonparametric regression problem

$$y_i = \mu(t_i) + \epsilon_i \quad (1.2)$$

where μ and t_i are as in (1.1) and ϵ_i are independent and identically distributed (i.i.d.) random variables with mean zero and finite variance. For this problem linear smoothers such as kernel regression estimators, smoothing splines, and local polynomial smoothers are known to provide consistent estimators of the regression function (Stone, 1977; Wahba and Wold, 1975; Fan, 1992). Good performance of these estimators depends critically on the choice of a smoothing parameter, which controls the variance versus bias trade-off of the estimator.

Performance of a nonparametric regression estimator is generally assessed pointwise by squared error loss of the form,

$$L(t, \lambda) = [\hat{\mu}_\lambda(t) - \mu(t)]^2,$$

where $\mu(t)$ is the true regression function and $\hat{\mu}_\lambda(t)$ is an estimator of the regression function depending on the smoothing parameter λ . Globally, average squared error (ASE) or its expectation (MASE)

$$\begin{aligned} ASE(\lambda) &= \frac{1}{n} \sum_{i=1}^n L(t_i, \lambda), \\ MASE(\lambda) &= E[ASE(\lambda)], \end{aligned}$$

are often used, where t_i are a selected set of points, generally the design points.

In the ordinary nonparametric regression problem (1.2) a number of bandwidth selectors based on minimizing an estimate of squared prediction risk have been devised. These include CV and Mallows' C_L and a number of similar estimates. Kernel regression estimators using the selected bandwidths have been shown to be consistent (Härdle and Marron, 1985; Li, 1987; Wong, 1983), although the rate of convergence of the selected bandwidths to the truly optimal bandwidth is known to be slow (Härdle, Hall and Marron, 1988). Li (1985, 1986) studied the consistency properties of cross-validated and generalized cross-validated estimators for smoothing splines. "Plug-in" estimators, such as those suggested for nonparametric density estimation (Jones, Marron and Park, 1991) have also been developed (Gasser, Kneip and Köhler, 1991; Hermann, Gasser and Kneip, 1992).

In this paper we consider the binary regression problem (1.1) using the kernel nonparametric regression estimator:

$$\hat{\mu}_\lambda(t) = \sum_{i=1}^n h_{\lambda i}(t) y_i$$

where $h_{\lambda i}(t)$ can represent any of the kernel weights:

- Priestley-Chao weights (Priestley and Chao, 1972; Benedetti, 1977)

$$h_{\lambda i}(t) = K\left(\frac{t - t_i}{\lambda}\right) K\left(\frac{t_i - t_{i-1}}{\lambda}\right);$$

- Nadaraya-Watson weights (Nadaraya, 1964 ; Watson, 1964)

$$h_{\lambda i}(t) = \frac{K\left(\frac{t - t_i}{\lambda}\right)}{\sum_j K\left(\frac{t - t_j}{\lambda}\right)};$$

- Gasser-Müller weights (Gasser and Müller, 1979)

$$h_{\lambda i}(t) = \int_{s_{i-1}}^{s_i} K\left(\frac{t-s}{\lambda}\right) ds,$$

with $s_0 = 0$, $s_n = 1$ and $s_i = (t_i + t_{i+1})/2$ for $1 \leq i < n$,

where $K(t)$ is a kernel function with properties described below.

We will most often need the kernel weights at the design points. For convenience of notation, we define $h_{\lambda ij} = h_{\lambda j}(t_i)$, with corresponding hat matrix $H(\lambda) = [h_{\lambda ij}]$.

Then

$$nASE(\lambda) = \| H(\lambda)y - \mu \|^2$$

and

$$nMASE(\lambda) = nE[ASE(\lambda)] = \text{tr} \Sigma H(\lambda) H(\lambda)' + \| H(\lambda)\mu - \mu \|^2$$

where μ denotes the vector $\mu_i = \mu(t_i)$, Σ is the diagonal matrix $\Sigma_{ii} = \sigma^2(t_i) = \mu(t_i)[1 - \mu(t_i)]$ and $\|x\|^2 = x'x$.

The usual assumptions about the kernel and regression functions used in the ordinary nonparametric regression problem will also be made here.

We will require the following condition on the spacing of the design points:

$$\text{A) } |t_i - t_{i-1} - \frac{1}{n}| = o(\frac{1}{n}).$$

This is required to assure the consistency of the kernel estimator for a deterministic sequence of bandwidths.

K is called a kernel of order p if the first $p - 1$ moments of K are 0 and the p^{th} moment

$$S_K = \int x^p K(x) dx$$

is not zero. Define

$$N_K = \int K^2(x) dx.$$

The kernel is assumed to have the following properties:

$$\text{B) } K \text{ is symmetric with support on the interval } (-\frac{1}{2}, \frac{1}{2}).$$

$$\text{C) } K \text{ is Lipschitz continuous of order } \alpha > 0.$$

The finite index set L_n , containing the λ 's will be assumed to have the following properties:

D) $L_n \subset (An^{-\frac{1}{2p+1}}, Bn^{-\frac{1}{2p+1}})$ where A and B are positive constants and p is the order of the kernel K .

E) The cardinality of L_n is $O(n^{m-2p/(1+2p)})$, where $m \geq 2$.

The following conditions for n sufficiently large are required on the smoother matrix $H(\lambda)$:

F) $|h_{\lambda ii}| \leq C/(n\lambda)$.

G) $\xi(\lambda)$, the largest eigenvalue of $H(\lambda)H'(\lambda)$ is bounded.

Finally, we require a smoothness condition on the regression function:

H) $\mu(t)$ has a square integrable p^{th} derivative.

Of course, since $\mu(t)$ is a probability for each t , $0 \leq \mu(t) \leq 1$.

Remark: It should be noted that the kernel estimator is inadmissible if $\xi(\lambda) > 1$ (Cohen, 1966). Although Conditions F and G would appear to be guaranteed by conditions A through C, this is not always evident. Clearly $|h_{\lambda ii}| < C/(n\lambda)$ for some C for the Gasser-Müller and Priestley-Chao weights; in fact, it can be shown that the Gasser-Müller weights satisfy $|h_{\lambda ii}| < C/n$. This fact implies that $C^{-1}H(\lambda)H'(\lambda)$ is substochastic for Gasser-Müller weights with positive ($p = 2$) kernels. Lemma 1 and Corollary 2 in the Appendix show the boundedness of the largest eigenvalue in this case. The Nadaraya-Watson weights can also be shown to satisfy $|h_{\lambda ij}| < C/n$ provided that $n^{1-\alpha}/\lambda \rightarrow 0$ for n sufficiently large or equivalently $\alpha > \frac{2p+1}{2p}$; in this case, condition G would also be satisfied for positive kernels. Conditions on the kernel function to yield condition G for Priestley-Chao weights are in general unknown.

Although here we restrict our discussion to nonparametric kernel regression, we conjecture that the consistency results on the choice of smoothing parameter remain true for many other linear smoothers, such as smoothing splines, local polynomial smoothers, or the many methods considered by Buja, Hastie and Tibshirani (1989).

In Section 2 we consider the consistency of CV and C_L for binary regression using adaptations of results for ordinary nonparametric regression. The methods of proof for convergence of $ASE(\lambda)$ were suggested by the results of Li (1987). The methods of proof for convergence of the selected bandwidth were suggested by the results of Härdle, Hall and Marron (1988) referred to as HHM88. A central limit theorem for degenerate U -statistics based on independent but not necessary identically distributed random variables (analogous to that of Hall, 1984, in the

i.i.d. case) is essential and such a theorem is proved here. Section 3 is an example. Section 4 gives simulation results related to the example.

Section 5 is a discussion of some methods which we find to be less satisfactory for bandwidth selection for heuristic reasons. These include weighted least squares where the weights depend on estimates of the variance function, and the “cross-validated likelihood” method of Hastie and Tibshirani (1990, p. 159) and Azzalini, Bowman and Härdle (1989).

2 Asymptotic Optimality of CV and C_L

We consider two bandwidth selection criteria based on $ASE(\lambda)$ and show that these lead to asymptotically optimal estimators of the regression function. The proofs in this section were inspired by Li (1989) and equation numbers of the form La.b refer to equation a.b in that paper.

Mallow’s C_L selects a bandwidth λ_{CL} which minimizes

$$C_L(\lambda) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{\mu}_\lambda(t_i)]^2 + \frac{2}{n} \sum_{i=1}^n h_{\lambda ii} \hat{\sigma}_\lambda^2(t_i) \quad (2.1)$$

where $\hat{\sigma}_\lambda^2(t_i) = \hat{\mu}_\lambda(t_i)[1 - \hat{\mu}_\lambda(t_i)]$. The CV criterion selects a bandwidth λ_{CV} which minimizes

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{\mu}_\lambda^{-i}(t_i)]^2 \quad (2.2)$$

where $\hat{\mu}_\lambda^{-i}(t_i)$ is the kernel regression estimator of $\mu(t_i)$ for the data set including all of the data except y_i . We will show that

$$\frac{ASE(\lambda_C)}{\min_{\lambda \in L_n} ASE(\lambda)} \xrightarrow{P} 1 \quad (2.3)$$

where λ_C is the bandwidth selected by one of the criteria, and that letting $\lambda_{opt} = \min_{\lambda \in L_n} ASE(\lambda)$ or $\lambda_{opt} = \min_{\lambda \in L_n} MASE(\lambda)$,

$$n^{3/10}(\lambda_c - \lambda_{opt}) \xrightarrow{D} N(0, \sigma_c^2) \quad (2.4)$$

where the value of σ_c^2 depends on which criterion is being optimized.

Proof of (2.3) follows by showing first the uniform convergence of loss to risk. For this, assumptions D and E concerning the discrete index set L_n of bandwidths play a fundamental

role. These assumptions are not too restrictive. We find that the pointwise asymptotic bias of the kernel estimator is the same as in the continuous case (Gasser and Müller, 1979):

$$Bias[\hat{\mu}_\lambda(t)] = (-1)^p \lambda^p \mu^{(p)}(t) S_K / p! + o(\lambda^p) \quad (2.5)$$

and the pointwise asymptotic variance is

$$Var[\hat{\mu}_\lambda(t)] = \frac{1}{n\lambda} \sigma^2(t) N_K + o(1/n\lambda). \quad (2.6)$$

where S_K and N_K are defined in assumption C. Therefore any sequence of bandwidths that leads to consistent estimators must satisfy Assumption D. Assumptions D and E also ensure that for $m \geq 2$, $n^{-m} \sum_{\lambda \in L_n} MASE(\lambda)^{-m} \rightarrow 0$ which is a key fact in the proofs. The Lipschitz continuity of $MASE(\lambda)$ as a function of λ ensures that the results extend to the entire interval $(An^{-\frac{1}{2p+1}}, Cn^{-\frac{1}{2p+1}})$.

We obtain the following convergence result (proof is in the Appendix):

Lemma 2: $\sup_{\lambda \in L_n} \left| \frac{ASE(\lambda)}{MASE(\lambda)} - 1 \right| \xrightarrow{P} 0$ as $n \rightarrow \infty$.

Once Lemma 2 is established, we need only show that:

$$\frac{ASE(\lambda_C) - \min_{\lambda} MASE(\lambda)}{\min_{\lambda} MASE(\lambda)} \xrightarrow{P} 0.$$

2.1 C_L

Theorem 1 *Mallow's C_L is asymptotically optimal.*

Proof: In matrix notation $nC_L(\lambda) = \| \mathbf{y} - H(\lambda)\mathbf{y} \|^2 + 2tr \hat{\Sigma}_\lambda H(\lambda)$ where $\hat{\Sigma}_\lambda$ is the diagonal matrix with $\hat{\Sigma}_{\lambda ii} = \hat{\sigma}_\lambda^2(t_i)$ and \mathbf{y} is the vector $(y_1 \cdots y_n)'$.

$$\begin{aligned} n[C_L(\lambda) - ASE(\lambda)] &= \| \mathbf{y} - \mu \|^2 + 2(\mathbf{y} - \mu)'[\mu - H(\lambda)\mu] \\ &\quad + 2tr \Sigma H(\lambda) - 2(\mathbf{y} - \mu)'H(\lambda)(\mathbf{y} - \mu) - 2tr(\Sigma - \hat{\Sigma}_\lambda)H(\lambda). \end{aligned}$$

Note that $\| \mathbf{y} - \mu \|^2$ does not depend on λ , and so is not important for minimization. The required result will follow if we can show that:

$$\sup_{\lambda \in L_n} \frac{(\mathbf{y} - \mu)'[\mu - H(\lambda)\mu]}{nMASE(\lambda)} \xrightarrow{P} 0, \quad (2.7)$$

$$\sup_{\lambda \in L_n} \frac{\text{tr} \Sigma H(\lambda) - (\mathbf{y} - \mu)' H(\lambda) (\mathbf{y} - \mu)}{n \text{MASE}(\lambda)} \xrightarrow{P} 0, \quad (2.8)$$

and

$$\sup_{\lambda \in L_n} \frac{\text{tr}(\Sigma - \hat{\Sigma}_\lambda) H(\lambda)}{n \text{MASE}} \xrightarrow{P} 0. \quad (2.9)$$

Throughout, D will denote a generic constant whose value will change with context.

Since the moments of a Bernoulli random variable are bounded by 1, (2.7) follows by Whittle's inequality for sums (Whittle, 1960) and Chebyshev's inequality, just as L2.2.

By Whittle's inequality for quadratic forms (Whittle, 1960)

$$E \left(|\text{tr} \Sigma H(\lambda) - (\mathbf{y} - \mu)' H(\lambda) (\mathbf{y} - \mu)|^{2m} \right) \leq D \left(\sum_{i=1}^n \sum_{j=1}^n h_{\lambda ij}^2 \tau_i^2 \tau_j^2 \right)^m,$$

where by Hölder's Inequality,

$$\tau_i = \left[E |y_i - \mu(t_i)|^{4m} \right]^{\frac{1}{4m}} \leq \sigma(t_i) \left[E |y_i - \mu(t_i)|^{4m-2} \right]^{1/4m-2} \leq \sigma(t_i);$$

because y_i is Bernoulli. So (2.8) follows by Chebyshev's inequality as in L2.3.

Finally, for n sufficiently large,

$$\begin{aligned} |\hat{\sigma}_\lambda^2(t) - \sigma^2(t)| &\leq |\hat{\mu}_\lambda(t) - \mu(t)| + |\hat{\mu}_\lambda(t) + \mu(t)| |\hat{\mu}_\lambda(t) - \mu(t)| \\ &\leq 4 |\hat{\mu}_\lambda(t) - \mu(t)|. \end{aligned}$$

So

$$\begin{aligned} P \left(\frac{\text{tr}(\Sigma - \hat{\Sigma}_\lambda) H(\lambda)}{n \text{MASE}} > \Delta \right) &\leq P \left(\frac{4 \sum_{i=1}^n h_{\lambda ii} |\hat{\mu}_\lambda(t_i) - \mu(t_i)|}{n \text{MASE}} > \Delta \right) \\ &\leq \frac{4^{2m} E \left(\sum_{i=1}^n |h_{\lambda ii}| |\hat{\mu}_\lambda(t_i) - \mu(t_i)| \right)^{2m}}{\Delta^{2m} n^{2m} \text{MASE}(\lambda)^{2m}}. \end{aligned}$$

By condition F, $|h_{\lambda ii}| \leq \frac{C}{n\lambda}$ so

$$\begin{aligned} E \left(\sum_{i=1}^n |h_{\lambda ii}| |\hat{\mu}_\lambda(t_i) - \mu(t_i)| \right)^{2m} &\leq \left(\frac{C}{n\lambda} \right)^{2m} \left[\sum_{i=1}^n [\hat{\mu}_\lambda(t_i) - \mu(t_i)]^2 \right]^m \\ &\leq \left(\frac{C}{n\lambda} \right)^{2m} n^m \text{MASE}(\lambda)^m \end{aligned}$$

and the desired result follows.

2.2 CV

The proof of asymptotic optimality rests on the fact that for large n , the difference between $\hat{\mu}_\lambda(t_i)$ and $\hat{\mu}_\lambda^{-i}(t_i)$ is small. Define $\hat{\mu}_\lambda^-$ to be the vector with i^{th} element $\hat{\mu}_\lambda^{-i}(t_i)$, and define $K(\lambda)$ to be the matrix such that $K(\lambda)\mathbf{y} = \hat{\mu}_\lambda^-$.

Theorem 2 *If $K(\lambda)$ is such that $\psi(\lambda)$ the largest eigenvalue of $K(\lambda)K'(\lambda)$ is bounded, then $CV(\lambda)$ is asymptotically optimal.*

Proof: Let $n\widetilde{ASE}(\lambda) = \|\mu - K(\lambda)\mathbf{y}\|^2$ and $M\widetilde{ASE}(\lambda) = E[\widetilde{ASE}(\lambda)]$. Just as in Lemma 2, we can readily show that

$$\sup_{\lambda \in L_n} \left| \frac{\widetilde{ASE}(\lambda)}{M\widetilde{ASE}(\lambda)} - 1 \right| \xrightarrow{P} 0 \quad (2.10)$$

Note that as the diagonal elements of $K(\lambda)$ are all zero, $tr\Sigma K(\lambda) = 0$. Following the steps of Theorem 1 we find that

$$n[CV(\lambda) - \widetilde{ASE}(\lambda)] = \|\mathbf{y} - \mu\|^2 + 2(\mathbf{y} - \mu)'(\mu - K(\lambda)\mu) + 2tr\Sigma K(\lambda) - 2(\mathbf{y} - \mu)'K(\lambda)(\mathbf{y} - \mu)$$

and that $\sup_{\lambda \in L_n} [(\mathbf{y} - \mu)'(\mu - K(\lambda)\mu)]/nM\widetilde{ASE}(\lambda) \xrightarrow{P} 0$ and $\sup_{\lambda \in L_n} [tr\Sigma K(\lambda) - (\mathbf{y} - \mu)'K(\lambda)(\mathbf{y} - \mu)]/nM\widetilde{ASE}(\lambda) \xrightarrow{P} 0$.

The proof is therefore complete if we can show that for any sequence $\{\lambda_n\}$ with $\lambda_n \in L_n$

$$\frac{M\widetilde{ASE}(\lambda_n)}{MASE(\lambda_n)} \rightarrow 1. \quad (2.11)$$

The proof of (2.11) is in the Appendix.

Remark: It is well-known (c.f. Milnor, 1963) that if corresponding entries of two symmetric matrices ($n \times n$) differ by at most ϵ then their corresponding eigenvalues differ by at most $n\epsilon$. When a positive kernel is used, this implies, for instance, in the case of the Gasser-Müller weights and in the case of the Nadaraya-Watson weights satisfying $\alpha > \frac{2p+1}{2p}$, that since $h_{\lambda ij}$ is bounded by C/n and the largest eigenvalue of $H(\lambda)H'(\lambda)$ is bounded, that the largest eigenvalue of $K(\lambda)K'(\lambda)$ is also bounded, since the corresponding entries of $H(\lambda)H'(\lambda)$ and $K(\lambda)K'(\lambda)$ differ by at most C/n .

2.3 Convergence of the Selected Bandwidth

HHM88 addressed the problem of smoothing parameter selection for the nonparametric curve estimators in the ordinary nonparametric regression problem (1.2). They proved a central limit theorem which quantifies a convergence rate of a class of automatically selected bandwidths to the “optimal bandwidth”, that is, the minimizer of the $ASE(\lambda)$ or $MASE(\lambda)$. Here we sketch the reasons why an analogous result holds for binary regression.

As remarked by HHM88 most bandwidth selectors are based on the minimization of some function of the residual sum of squares

$$RSS(\lambda) = \frac{1}{n} \sum (y_i - \mu_\lambda(t_i))^2.$$

Here we have considered $CV(\lambda)$ and $C_L(\lambda)$. As shown in the appendix of HHM88,

$$\frac{CV(\lambda)}{RSS(\lambda)} = 1 + \frac{2}{n\lambda} K(0) + O\left(\frac{1}{n^2\lambda^2}\right). \quad (2.12)$$

A similar argument holds for $C_L(\lambda)$. A much larger list of bandwidth selectors satisfying (2.12) is given in HHM88. Essentially, we are interested in showing that a bandwidth selector λ_C minimizing a selection criterion $C(\lambda)$ which satisfies (2.12) converges to λ_{ASE} the bandwidth minimizing $ASE(\lambda)$ and λ_{MASE} the bandwidth minimizing $MASE(\lambda)$.

Similar to the result of HHM88 for the continuous case, we have the following results for binary regression:

Theorem 3 *Under assumptions A - G, when the kernel has order $p = 2$ and Hölder continuous second derivative on its support, and the regression function $\mu(t)$ has a uniformly continuous integrable second derivative then:*

$$\begin{aligned} n^{3/10}(\lambda_C - \lambda_{ASE}) &\xrightarrow{D} N(0, \sigma_1^2) \\ n[ASE(\lambda_C) - ASE(\lambda_{ASE})] &\xrightarrow{D} D_1 \chi_1^2 \end{aligned}$$

and also

$$\begin{aligned} n^{3/10}(\lambda_{MASE} - \lambda_{ASE}) &\xrightarrow{D} N(0, \sigma_2^2) \\ n(ASE(\lambda_{MASE}) - ASE(\lambda_{ASE})) &\xrightarrow{D} D_2 \chi_1^2 \end{aligned}$$

where D_i and σ_i^2 are defined as part of the proof in the Appendix.

3 Analysis of the Periparturient Recumbent Cows Data

Clark et al (1987) collected a set of biochemical and haematological measures on dairy cows suffering from periparturient recumbency, a common problem often leading to death or euthanasia. The purpose of the study was to evaluate these measures as predictors of recovery.

For the purposes of this analysis, 110 cows were classified as either recovered or not recovered. (Cows that did not recover died, either from the condition or from euthanasia by the farmer or veterinarian due to poor prognosis). Although a number of measures were collected, the focus here is on a single variable, serum urea. In cattle, increased serum urea may be due to a number of causes such as shock, increased protein catabolism and/or kidney damage. Both high and low values are considered indicators of poor health.

Figure 1 shows the raw data and fits for Nadaraya-Watson and Gasser-Müller kernel estimators using quadratic weights and bandwidths selected by either CV or C_L . The raw data, indicated by asterisks on the plot, are the percentage of cows recovering at each value of serum urea. Due to round-off, as many as 3 cows were measured at some level of urea, but most levels had only 1 animal, leading to proportions of 0, 1/3, 1/2, 2/3 or 1.

CV selected a bandwidth of .51 for the Gasser-Müller kernel, and .11 for the Nadaraya-Watson kernel, leading to undersmoothing of the latter. (See Altman, 1992 for an assessment of the fit of the smooth using the method of Azzalini, Bowman and Härdle, 1989.) Conversely, C_L undersmoothed using the Gasser-Müller kernel ($\lambda = .19$) and produced a smoother fit with the Nadaraya-Watson kernel ($\lambda = .52$).

These results indicate that the sample sizes required for automatic bandwidth selection in binary regression are somewhat larger than those needed for continuous data. A simulation study which explores this idea in more detail is discussed in Section 4.

4 Simulations

In order to understand the results of the periparturient recumbent cows example, we undertook a small simulation study. Design points were simulated from the mixture of normals

$$.25N(1.4, 0.4) + .60N(2.2, 0.4) + .15N(3.0, 0.2)$$

truncated to the interval $[0.7, 3.4]$. This mixture approximates a kernel density estimate of the design points for the data set. The average curve was defined by

$$\mu(t) = 19.1 - 57.1t + 63.0t^2 - 31.9t^3 + 7.6t^4 - 0.69t^5$$

which is the 5th degree polynomial fit to the periparturient recumbent cows data slightly adjusted to ensure positivity on the region of definition. Figure 2 shows $\mu(t)$ and the fit of the Gasser-Muller kernel with $\lambda = .51$. 100 Bernoulli response “curves” were generated for sample sizes of 100, 500 and 1000 design points. To reduce computation time, design points were generated just once for each sample size.

Nadaraya-Watson and Gasser-Müller kernel regression estimators were fitted to each generated curve, using quadratic kernels. Bandwidths for each curve were selected using one of three criteria: average squared error, CV and C_L .

Figure 3 shows the difference between the logarithms of the selected and optimal bandwidths for each combination of kernel and selector for samples sizes of 100, 500 and 1000. As the sample size increases, the distribution of differences slowly becomes less spread out, although there is considerable variability even at $n = 1000$. There is no marked trend towards over or undersmoothing nor to either selector dominating.

Figure 4 shows the ratio $ASE(\lambda_C)/\min_{\lambda} ASE(\lambda)$ for each combination of kernel and selector for the same realizations depicted in Figure 2. The relative ASE does approach 1 as the sample size increases, but convergence is slow. There is considerable variability even at $n = 1000$.

5 Other Methods of Bandwidth Selection

Because the variance of y_i depends on its mean in the binary regression problem, an average weighted squared error criterion of the form

$$AWSE(\lambda) = \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_{\lambda}(t_i) - \mu(t_i)]^2 w(t_i) \quad (5.1)$$

with weights depending on the local variance may be preferable to the use of $ASE(\lambda)$ for bandwidth selection. Although we favor this idea on philosophical grounds, for reasons discussed below, we were unable to find an appropriate way to determine the weights.

For Bernoulli data, $\sigma^2(t) = 0$ whenever $\mu(t)$ is 0 or 1. Equation (2.5) shows that, if there is curvature at such points, the asymptotic bias of the regression estimator is non-zero. Hence for general mean functions $\mu(t)$, the asymptotic weighted mean squared error would blow up where local extrema reach 0 or 1, if the weights are chosen to be inversely proportional to the variance.

However, even if the mean function is in some region $0 < \delta_1 \leq \mu(t) \leq \delta_2 < 1$ the use of weighted least squares is problematic, due to the need to estimate the weights from the data. The appropriate adjustments to CV and C_L for the weights are

$$WCV(\lambda) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{\mu}_\lambda^{-i}(t_i)]^2 w(t_i)$$

and

$$WCL(\lambda) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{\mu}_\lambda(t_i)]^2 w(t_i) + \frac{2}{n} \sum_{i=1}^n h_{\lambda ii} \hat{\sigma}_\lambda^2 w(t_i)$$

respectively, which are estimators of

$$AWSE(\lambda) = \frac{1}{n} \sum_{i=1}^n [y_i^* - \hat{\mu}_\lambda(t_i)]^2 w(t_i)$$

To understand the problem, note that by (2.5) and (2.6), $\hat{\mu}_\lambda(t_i) = \mu(t_i) + (-1)^p \frac{\lambda^p}{p!} S_K \mu^{(p)}(t_i) + O_p((n\lambda)^{-1/2}) + o_p(\lambda^p)$ and therefore

$$\hat{\sigma}_\lambda^2(t) = \sigma^2(t) \left[1 + (-1)^p \frac{\lambda^p}{p!} \frac{S_K}{\sigma^2(t)} \mu^{(p)}(t) [1 - 2\mu(t)] + \frac{1 - 2\mu(t)}{\sigma^2(t) \sqrt{n\lambda}} O_p(1) + o_p(\lambda^p) \right]. \quad (5.2)$$

We would like to use $w(t_i) = 1/\sigma^2(t_i)$, but instead use $\hat{w}(t_i) = 1/\hat{\sigma}_\lambda^2(t_i)$. We then find that

$$\begin{aligned} E \left\{ [y_i^* - \hat{\mu}_\lambda(t_i)]^2 \hat{w}(t_i) \right\} &= E \left\{ [y_i^* - \hat{\mu}_\lambda(t_i)]^2 w(t_i) \right\} \\ &\quad + [y_i^* - \mu(t_i)]^2 \left\{ (-1)^p \frac{\lambda^p}{p!} \frac{S_K}{\sigma^2(t)} \mu^{(p)}(t) [1 - 2\mu(t)] + O_p(1/\sqrt{n\lambda}) \right\} \\ &\quad + o_p(\lambda^p). \end{aligned}$$

The second term in this expansion is larger than $MWASE(\lambda)$ which is $O(\lambda^{2p}) + O(1/n\lambda)$. The approximations in $WCL(\lambda)$ and $WCV(\lambda)$ do not correct for this error term. We have been unable to find a method of estimating weights that does not suffer from this problem.

The bandwidth selection methods explored in this paper have been based on quadratic loss. In analogy with maximum likelihood methods in parametric problems, the use of modified maximum likelihood has been suggested for bandwidth selection (Azzalini, Bowman and Härdle, 1989; Hastie and Tibshirani, 1990, p.159). The bandwidth is selected to minimize the leave-one-out estimator of the log-likelihood:

$$ML(\lambda) = \frac{1}{n} \sum_{i=1}^n \log(\hat{\mu}_{\lambda}^{-i}(t_i)^{y_i} [1 - \hat{\mu}_{\lambda}^{-i}(t_i)]^{1-y_i})$$

where $\hat{\mu}_{\lambda}^{-i}(t_i)$ is defined following (2, Section 2).

The $ML(\lambda)$ criterion was first proposed by Habbema, Hermans and Van den Broek (1974) and Duin (1976) on heuristic grounds for bandwidth selection for nonparametric density estimation. Bowman (1980, 1984) showed that for density estimation $ML(\lambda)$ is a cross-validation criterion based on Kullback-Leibler loss for discrete density estimation problems and can be viewed in that context for continuous density estimation problems. However, Schuster and Gregory (1981) showed that $ML(\lambda)$ is not asymptotically optimal for density estimation of continuous distributions with long tails.

Using the argument of Bowman (1984), it is easy to derive a loss function for binary regression for which $ML(\lambda)$ is a leave-one-out estimator. This has the form:

$$nL(\mu, \hat{\mu}_{\lambda}) = \sum_{i=1}^n \mu_{\lambda}(t_i) \log[\mu(t_i)/\hat{\mu}_{\lambda}(t_i)] + \sum_{i=1}^n (1 - \mu_{\lambda}(t_i)) \log[(1 - \mu(t_i))/(1 - \hat{\mu}_{\lambda}(t_i))]$$

which is clearly not Kullback-Leibler loss. There do not appear to be any compelling reasons to consider this as an appropriate loss function for this problem.

6 Appendix

An $n \times n$ square matrix $P = [p_{ij}]$ is said to be substochastic if it consists of non-negative elements satisfying $\sum_{j=1}^n p_{ij} = p_i \leq 1$.

Lemma 1 : For any vector \mathbf{x} , if P is substochastic then $\|P\mathbf{x}\|^2 \leq \|\mathbf{x}\|^2$.

Proof:

$$\|P\mathbf{x}\|^2 = \sum_{i=1}^n \left| \sum_{j=1}^n p_{ij} x_j \right|^2$$

$$\begin{aligned}
&= \sum_{i=1}^n p_{i.}^2 \left| \sum_{j=1}^n \frac{p_{ij}}{p_{i.}} x_j \right|^2 \\
&\leq \sum_{i=1}^n p_{i.}^2 \sum_{j=1}^n \frac{p_{ij}}{p_{i.}} |x_j|^2 \quad \text{by convexity} \\
&\leq \sum_{i=1}^n \sum_{j=1}^n p_{ij} |x_j|^2 \\
&\leq \sum_{j=1}^n |x_j|^2.
\end{aligned}$$

In particular, we have the following corollary:

Corollary 1: If P is substochastic and symmetric, then the largest eigenvalue is bounded by 1.

Lemma 2 :

$$\sup_{\lambda \in L_n} \left| \frac{ASE(\lambda)}{MASE(\lambda)} - 1 \right| \xrightarrow{P} 0.$$

Proof:

$$n[ASE(\lambda) - MASE(\lambda)] = 2(\mu - H(\lambda)\mu)'H(\lambda)(\mathbf{y} - \mu) + \|H(\lambda)(\mathbf{y} - \mu)\|^2 - \text{tr}\Sigma H(\lambda)H(\lambda)'$$

By Whittle's Inequality for sums,

$$\begin{aligned}
E \left\{ [(\mathbf{y} - \mu)'H(\lambda)'(\mu - H(\lambda)\mu)]^{2m} \right\} &\leq D \|H(\lambda)'(\mu - H(\lambda)\mu)\|^{2m} \\
&\leq D\xi(\lambda)^m \|\mu - H(\lambda)\mu\|^{2m} \\
&\leq D\xi(\lambda)^m n^m MASE(\lambda)^m,
\end{aligned}$$

so

$$\sup_{\lambda \in L_n} \frac{(\mathbf{y} - \mu)'H(\lambda)'(\mu - H(\lambda)\mu)}{nMASE(\lambda)} \xrightarrow{P} 0.$$

By Whittle's Inequality for quadratic forms and the proof of (2.9),

$$\begin{aligned}
E | \text{tr}\Sigma H(\lambda)H(\lambda)' - (\mathbf{y} - \mu)'H(\lambda)'H(\lambda)(\mathbf{y} - \mu) |^{2m} &\leq D(\text{tr}\Sigma H(\lambda)H(\lambda)'H(\lambda)H(\lambda)'\Sigma)^m \\
&\leq [\frac{1}{4}\xi(\lambda)]^m D \text{tr}H(\lambda)H(\lambda)\Sigma \\
&\leq [\frac{1}{4}\xi(\lambda)]^m D n^m MASE(\lambda)^m,
\end{aligned}$$

so

$$\sup_{\lambda \in L_n} \frac{\text{tr} \Sigma H(\lambda) H(\lambda)' - (\mathbf{y} - \mu)' H(\lambda)' H(\lambda) (\mathbf{y} - \mu)}{n MASE(\lambda)} \xrightarrow{P} 0.$$

Proof of (2.11): For any sequence $\{\lambda_n\}$ with $\lambda_n \in L_n$

$$\frac{\widetilde{MASE}(\lambda_n)}{MASE(\lambda_n)} \rightarrow 1.$$

Proof proceeds separately for the Priestley-Chao, Nadaraya-Watson and Gasser-Müller weights.

Nadaraya-Watson Weights:

$$\hat{\mu}_{\lambda}^{-i}(t_i) - \mu(t_i) = \frac{\hat{\mu}_{\lambda}(t_i) - \mu(t_i) - h_{\lambda ii}[y_i - \mu(t_i)]}{1 - h_{\lambda ii}},$$

so:

$$\begin{aligned} E \left[\hat{\mu}_{\lambda}^{-i}(t_i) - \mu(t_i) \right]^2 &= \frac{E \left[\hat{\mu}_{\lambda}(t_i) - \mu(t_i) \right]^2 + h_{\lambda ii}^2 \sigma^2(t_i) - 2h_{\lambda ii} \text{Cov}(\hat{\mu}_{\lambda}(t_i), y_i)}{(1 - h_{\lambda ii})^2} \\ &= \frac{E \left[\hat{\mu}_{\lambda}(t_i) - \mu(t_i) \right]^2 - h_{\lambda ii}^2 \sigma^2(t_i)}{(1 - h_{\lambda ii})^2}. \end{aligned}$$

Therefore:

$$\begin{aligned} & \left| \frac{\widetilde{MASE}(\lambda) - MASE(\lambda)}{MASE(\lambda)} \right| \\ &= \frac{n MASE(\lambda) + \sum_{i=1}^n \sigma^2(t_i) (h_{\lambda ii}^2 - 2h_{\lambda ii}) - n MASE(\lambda) + (2h_{\lambda ii} - h_{\lambda ii}^2) n MASE(\lambda)}{(1 - h_{\lambda ii})^2 n MASE(\lambda)} \\ &\leq \frac{2n MASE(\lambda) (2|h_{\lambda ii}| + h_{\lambda ii}^2)}{(1 - h_{\lambda ii})^2 n MASE(\lambda)} \\ &\leq \frac{4B}{n\lambda} + o\left(\frac{1}{n\lambda}\right). \end{aligned}$$

Gasser-Müller Weights:

$$\hat{\mu}_{\lambda}^{-i}(t_i) = \hat{\mu}_{\lambda}(t_i) + (y_{i-1} - y_i) h_{\lambda ii},$$

so:

$$n \widetilde{MASE}(\lambda) = n MASE(\lambda)$$

$$\begin{aligned}
& + \sum_{i=1}^n E(y_{i-1} - y_i)^2 h_{\lambda ii}^2 \\
& + 2 \sum_{i=1}^n h_{\lambda ii} E[y_{i-1} - y_i][\hat{\mu}_\lambda(t_i) - \mu(t_i)].
\end{aligned}$$

Now $|y_{i-1} - y_i| \leq 1$ so

$$\begin{aligned}
\sum_{i=1}^n E(y_{i-1} - y_i)^2 h_{\lambda ii}^2 & \leq \sum_{i=1}^n h_{\lambda ii}^2 \\
& \leq \left(\frac{B}{n\lambda} \right)^2 \\
& = o(nMASE(\lambda)).
\end{aligned}$$

By condition F and since the bias of $\hat{\mu}_\lambda(t_i)$ is of the order of λ^p ,

$$\begin{aligned}
\sum_{i=1}^n h_{\lambda ii} E[y_{i-1} - y_i][\hat{\mu}_\lambda(t_i) - \mu(t_i)] & \leq \frac{B}{n\lambda} \sum_{i=1}^n |E[\hat{\mu}_\lambda(t_i) - \mu(t_i)]| \\
& = O(\lambda^{p-1}) \\
& = o(nMASE(\lambda)).
\end{aligned}$$

Priestley-Chao Weights

$$\hat{\mu}_\lambda^{-i}(t_i) = \hat{\mu}_\lambda(t_i) + \left(\frac{t_i - t_{i-1}}{\lambda} \right) \left[K \left(\frac{t_i - t_{i-1}}{\lambda} \right) y_{i+1} - K(0)y_i \right].$$

Proof of the result is similar to the proof for the Gasser-Müller weights and is left to reader.

Proof of Theorem 3:

Here the method of proof follows that in HHM88. The sequence of steps outlined in the Appendix there is meticulously followed here with the appropriate changes for binary regression. For the sake of completeness, the two main differences are given below. The first lemma they used is listed here and a proof is provided which differs slightly from theirs in order to accommodate Bernoulli random variables instead of mean zero i.i.d. errors. It is also necessary to develop a central limit theorem for degenerate U-statistics based on independent but not necessarily identically distributed random variables. Theorem 4 given below states the needed result, and an outline of its proof is provided. The proofs of the remaining results follow analogously to HHM88, and are omitted here.

As the constants in our Theorem 3 differ from those in HHM88 we give them first.

Let

$$C_0 = \left[N_k \int \sigma^2 \right] \left[S_k^2 \left(\int (\mu'')^2 \right) \right]^{-1/5},$$

and

$$C_1 = 2C_0^{-3} N_k \int \sigma^2 + 3C_0^2 S_k^2 \left(\int (\mu'')^2 \right).$$

Now letting $(*)$ denote convolution, define

$$\sigma_3^2 = 8C_0^{-3} \left(\int \sigma^4 \right) \left(\int (K * K - K * L)^2 \right) + 4C_0^3 S_k^2 \left(\int \sigma^2 \right) \left(\int (\mu'')^2 \right)$$

$$\sigma_3^2 = 8C_0^{-3} \left(\int \sigma^4 \right) \left(\int (K - L)^2 \right) + 4C_0^3 S_k^2 \left(\int \sigma^2 \right) \left(\int (\mu'')^2 \right)$$

The constants in Theorem 3 can now be defined as

$$\sigma_1^2 = C_1^{-2} \sigma_2^2,$$

$$\sigma_2^2 = C_1^{-2} \sigma_3^2,$$

$$D_1 = C_1 \frac{\sigma_1^2}{2}$$

$$D_2 = C_1 \frac{\sigma_2^2}{2}.$$

As in HHM88, the following two expansions play an important role in the proofs.

$$\begin{aligned} ASE'(\lambda_{ASE}) &= 0 \\ &= MASE'(\lambda_{ASE}) + \Delta'(\lambda_{ASE}) \\ &= (\lambda_{ASE} - \lambda_{MASE}) MASE(\lambda^*) + \Delta'(\lambda_{ASE}) \end{aligned} \tag{6.1}$$

where the prime ($'$) denotes derivative, λ^* is between λ_{ASE} and λ_{MASE} and $\Delta(\lambda) = ASE(\lambda) - MASE(\lambda)$.

$$C(\lambda) = \left[ASE(\lambda) + \hat{\sigma}^2 + \delta_1(\lambda) \right] \left[1 + \frac{2}{n\lambda} K(0) + O_p(n^{-2}\lambda^{-2}) \right] \tag{6.2}$$

where $n\hat{\sigma}^2 = \sum_{i=1}^n [y_i - \mu(t_i)]^2$ and $n\delta_1(\lambda) = 2 \sum_{i=1}^n [\hat{\mu}_\lambda(t_i) - \mu(t_i)][\mu(t_i) - y_i]$.

Lemma 3 (HHM88 Lemma 1):

For $m = 1, 2, \dots$ there is a constant D_3 such that

$$\sup_{\lambda \in L_n} E \left| \left(\frac{\lambda^{1/2}}{n^{-1}\lambda^{-1} + \lambda^4} \right) \Delta'(\lambda) \right|^{2m} \leq D_3 \quad (6.3)$$

and

$$\sup_{\lambda \in L_n} E \left| \left(\frac{\lambda^{1/2}}{n^{-1}\lambda^{-1} + \lambda^4} \right) \delta'_1(\lambda) \right|^{2m} \leq D_3. \quad (6.4)$$

Furthermore, there is an $\eta_1 > 0$ and a constant D_4 such that

$$E \left[\left(\frac{\lambda^{1/2}}{n^{-1}\lambda^{-1} + \lambda^4} \right) (\Delta'(\lambda) - \Delta'(\lambda_0)) \right]^{2m} \leq D_4 \left(\frac{\lambda - \lambda_0}{\lambda} \right)^{\eta_1 m} \quad (6.5)$$

and

$$E \left[\left(\frac{\lambda^{1/2}}{n^{-1}\lambda^{-1} + \lambda^4} \right) \delta'_1(\lambda) - \delta'_1(\lambda_0) \right]^{2m} \leq D_4 \left(\frac{|\lambda - \lambda_0|}{\lambda} \right)^{\eta_1 m} \quad (6.6)$$

whenever $\lambda, \lambda_0 \in L_n$ with $\lambda \leq \lambda_0$ and $\left| \frac{\lambda - \lambda_0}{\lambda} \right| \leq 1$.

Proof:

It should be noted that $\Delta_1(\lambda) = -\frac{\lambda}{2}\Delta'(\lambda)$ can be expanded into

$$\Delta_1(\lambda) = \Delta_{11}(\lambda) - \Delta_{12}(\lambda) + \Delta_{21}(\lambda) - \Delta_{22}(\lambda) + \Delta_{31}(\lambda) - \Delta_{32}(\lambda)$$

where

$$\begin{aligned} \Delta_{11}(\lambda) &= n^{-2} \sum_{i \neq j} \left[\frac{1}{n} \sum_{k=1}^n \frac{1}{\lambda^2} K \left(\frac{t_k - t_i}{\lambda} \right) K \left(\frac{t_k - t_j}{\lambda} \right) \right] [y_i - \mu(t_i)] [y_j - \mu(t_j)]. \\ \Delta_{12}(\lambda) &= n^{-2} \sum_{i < j} \left[\frac{1}{n} \sum_{k=1}^n \frac{1}{\lambda^2} \left(\frac{t_k - t_j}{\lambda} \right) K \left(\frac{t_k - t_i}{\lambda} \right) K' \left(\frac{t_k - t_j}{\lambda} \right) + \left(\frac{t_k - t_i}{\lambda} \right) K \left(\frac{t_k - t_j}{\lambda} \right) K' \left(\frac{t_k - t_i}{\lambda} \right) \right] \\ &\quad \times [y_i - \mu(t_i)] [y_j - \mu(t_j)]. \\ \Delta_{21}(\lambda) &= \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{n} \sum_{k=1}^n \frac{1}{\lambda} K \left(\frac{t_k - t_i}{\lambda} \right) \left(\frac{2}{n} \sum_{j=1}^n \frac{1}{\lambda} K \left(\frac{t_k - t_j}{\lambda} \right) \mu(t_j) - \mu(t_k) \right) - \right. \\ &\quad \left. \left(\frac{1}{n} \sum_{j=1}^n \frac{1}{\lambda} K' \left(\frac{t_k - t_j}{\lambda} \right) \mu(t_j) - \mu(t_k) \right) \right] [y_i - \mu(t_i)]. \\ \Delta_{22}(\lambda) &= \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{n} \sum_{k=1}^n \frac{1}{\lambda} K' \left(\frac{t_k - t_i}{\lambda} \right) \left(\frac{1}{n} \sum_{j=1}^n \frac{1}{\lambda} K \left(\frac{t_k - t_j}{\lambda} \right) \mu(t_j) - \mu(t_k) \right) [y_i - \mu(t_i)] \right]. \end{aligned}$$

$$\begin{aligned}\Delta_{31}(\lambda) &= n^{-2} \sum_{i=1}^n \left[\frac{1}{n} \sum_{k=1}^n \left(\frac{1}{\lambda^2} K \left(\frac{t_k - t_i}{\lambda} \right) \right)^2 \right] [(y_i - \mu(t_i))^2 - \sigma^2(t_i)] . \\ \Delta_{32}(\lambda) &= n^{-2} \sum_{i=1}^n \left[\frac{1}{n} \sum_{k=1}^n \left(\frac{1}{\lambda^2} \right) K \left(\frac{t_k - t_i}{\lambda} \right) K' \left(\frac{t_k - t_i}{\lambda} \right) \right] [(y_i - \mu(t_i))^2 - \sigma^2(t_i)] .\end{aligned}$$

In order to prove (6.5), for example, it suffices to consider each of the above six terms separately. By standard arguments involving the compactness of the support of K we have that

$$\left| \frac{1}{n} \sum_{k=1}^n \frac{1}{\lambda^2} K \left(\frac{t_k - t_i}{\lambda} \right) K \left(\frac{t_k - t_j}{\lambda} \right) - \frac{1}{n} \sum_{k=1}^n \frac{1}{\lambda_0^2} K \left(\frac{t_k - t_i}{\lambda_0} \right) K \left(\frac{t_k - t_j}{\lambda_0} \right) \right| \leq D \lambda^{-1} \left| \frac{\lambda - \lambda_0}{\lambda} \right| .$$

The result now follows by taking expectations of each term separately and applying Theorem 2 of Whittle (1960).

The proofs of the other inequalities are similar.

Hall (1984) used martingale theory to obtain a central limit theorem for degenerate U -Statistics with variable kernels and applied it to derive central limit theorems for the integrated square error of multivariate non-parametric density estimators. This limit theorem was also essential for the results in Hall and Marron (1987) and in HHM88. Here, we use a version of a central limit theorem due to Hoeffding (1948) for U -Statistics based on independent but not necessarily identically distributed random variables to derive an analogous result useful in binary regression.

Theorem 4 *Let*

$$U_n = \sum_{i,j=1}^n H_n(Y_i, Y_j),$$

where H_n is a symmetric function and Y_1, \dots, Y_n are independent non-identically distributed random variables (or vectors) and H satisfies for each $i \neq j$, $1 \leq i, j \leq n$:

1. $E[H_n(Y_i, Y_j)] = 0$ and $E[H_n(Y_i, Y_j) \mid Y_i] = 0$ a.e.;
2. and if $b_n = \max_{1 \leq i, j \leq n} E[H_n^2(Y_i, Y_j)]$ then $\sup_n(b_n) < \infty$.

Now let

$$\sigma^2(U_n) = \sum_{\substack{1 \leq i < j, k \leq n \\ i \neq j \neq k}} E[H_n(Y_i, Y_j)H_n(Y_i, Y_k)]; \quad (6.7)$$

and for each $i = 1, \dots, n$ let

$$X_{n,i} = \sum_{j=1}^n H_n(Y_i, Y_j)$$

and

$$G_{n,i}(x, y) = E[H_n(Y_i, x)H_n(Y_i, y)].$$

Now if

$$\left\{ \sum_{i=1}^n \sum_{j=1}^n E(G_{n,i}^2(Y_i, Y_j)) + \sum_{i=2}^n E(X_{n,i}^4) \right\} / \sigma^4(U_n) \rightarrow 0 \quad (6.8)$$

then U_n is asymptotically normally distributed with mean 0 and variance given by $\sigma^2(U_n)$.

Proof. It follows from Hoeffding (1948. p. 300) that $\sigma^2(U_n)$ is given by equation (6.7). A straightforward application of Brown's martingale central limit theorem (Brown 1971; Hall and Heyde 1980) as used by Hall (1984) in the case of i.i.d. random variables yields the result.

More explicitly, let

$$v_{nl} = E[X_{nl}^2 | Y_1, \dots, Y_{l-1}] \quad \text{and} \quad V_n^2 = \sum_{l=2}^n v_{nl}.$$

Clearly,

$$v_{ni} = \sum_{j=1}^{i-1} \sum_{k=1}^{i-1} G_n(Y_j, Y_k) = 2 \sum_{1 \leq j < k \leq n} G_n(Y_j, Y_k) + \sum_{j=1}^n G_n(Y_j, Y_j).$$

Now $E[G_n(Y_j, Y_k)G_n(Y_p, Y_r)] = 0$ unless $j = k = p = r$ or $j = k \neq p = r$ or $j = p$ and $k = r$.

Thus, if $l \leq m$

$$\begin{aligned} E(v_{nl}v_{nm}) &= 4 \sum_{1 \leq j < k \leq n} E[G_n^2(Y_j, Y_k)] + \sum_{j=1}^{l-1} \sum_{p=1}^{m-1} E[G_n(X_j, X_j)]E[G_n(X_p, X_p)] \\ &\quad + \sum_{r=1}^m \text{var}G_n(X_r, X_r). \end{aligned}$$

Clearly, if $s_n^2 = \sigma^2(U_n)$, condition (6.8) implies that

$$\frac{1}{s_n^4} E(V_n^2 - s_n^2)^2 \rightarrow 0. \quad (6.9)$$

Also,

$$E(X_{n,1}^2) = \sum_{j=1}^{i-1} E[H_n^2(Y_i, Y_j)]$$

and

$$\sigma^2(U_n) = s_n^2 = \sum_{i=2}^n \sum_{j=1}^{i-1} E[H_n^2(Y_i, Y_j)]$$

and

$$E[X_{n,i}^4] = \sum_{j=1}^{i-1} E[H_n^4(Y_i, Y_j)] + 3 \sum_{\substack{j,k=1 \\ j \neq k}}^{i-1} E[H_n^2(Y_i, Y_j)H_n^2(Y_i, Y_k)]$$

Hence it follows from condition (6.8) that

$$\frac{1}{s_n^4} \sum_{i=2}^n E(X_{n,i}^4) \rightarrow 0. \quad (6.10)$$

Now conditions (6.9) and (6.10) suffice for the martingale central limit theorem of Brown (1971) to hold (*cf.* Hall and Heyde (1980), Hall (1984)).

References

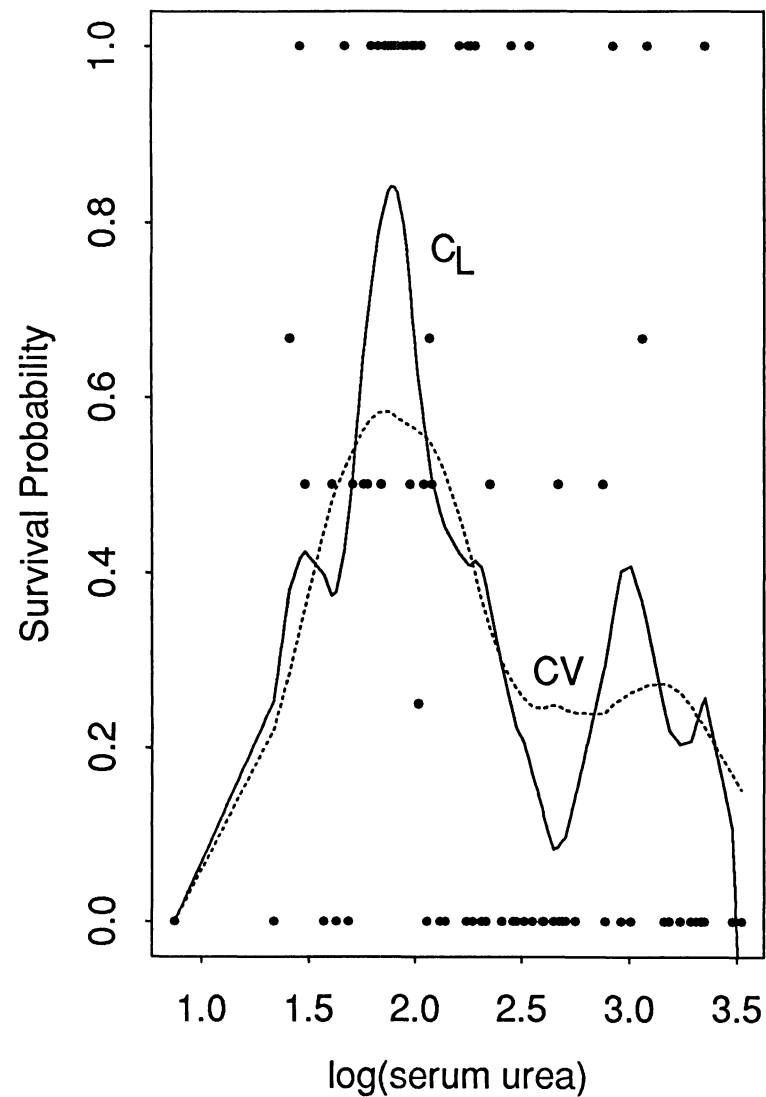
- Allen, D. M. (1974) "The Relationship between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, **16**, 1307-1325.
- Altman, N. S. (1992) "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." *The American Statistician*, **46**, 175-185.
- Azzalini, A., Bowman, A. W., and Härdle, W. (1989) "On the Use of Nonparametric Regression for Model Checking," *Biometrika*, **76**, 1-11.
- Benedetti, J. K. (1977) "On the Nonparametric Estimation of Regression Functions," *Journal of the Royal Statistical Society Series B*, **39**, 248-253.
- Bowman, A. W. (1980) "A Note of the Consistency of the Kernel Method for the Analysis of Categorical Data," *Biometrika*, **67**, 682-684.
- Bowman, A. W. (1984) "An Alternative Method of Cross-validation for the Smoothing of Density Estimates," *Biometrika*, **71**, 353-360.
- Brown, B. M. (1971). "Martingale Central Limit Theorems," *Annals of Mathematical Statistics*, **42**, 59-66.

- Buja, A., Hastie, T. and Tibshirani, R. (1989) "Linear Smoothers and Additive Models," (with discussion), *Annals of Statistics*, **17**, 453-555.
- Clark, R. G., Henderson, H. V., Hoggard, G. K., Ellison, R. S. and Young, B. J. (1987) "The Ability of Biochemical and Haematological Tests to Predict Recovery in Periparturient Recumbent Cows," *New Zealand Veterinary Journal*, 126-133.
- Cohen (1966) "All Admissible Linear Estimates of the Mean Vector." *Annals of Mathematical Statistics*, **37**, 458-463.
- Copas, J. B. (1983) "Plotting p Against x." *Applied Statistics*, **32**, 25-31.
- Duin, R. P. W. . (1976) "On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions," *I.E.E.E. Trans. Comput.* **C-25**, 1175-1179. (1976)
- Fan, J. (1992) "Design-adaptive Nonparametric Regression." *Journal of the American Statistical Association*, **87**, 998-1004.
- Fienberg, S. E. and Gong, G. D., (1984) Contribution to the discussion of a paper by J. M. Landwehr, D. Pregibon and A. C. Shoemaker. *Journal of the American Statistical Association*, **79**, 72-77.
- Fowlkes, E. B. (1987) "Some Diagnostics for Binary Logistic Regression via Smoothing." *Biometrika*, **74**, 503-515.
- Gasser, T., Kneip, A., Köhler, W. (1991) "A Flexible and Fast Method for Automatic Smoothing," *Journal of the American Statistical Association*, **86**, 643-652.
- Gasser, T., Müller, H. G. (1979) "Kernel Estimation of Regression Functions," in *Smoothing Techniques for Curve Estimation* ed. Gasser T. and Rosenblatt, M., 23-67, Lecture Notes in Mathematics 757, Springer-Verlag, Heidelberg.
- Geisser, S. (1975) "The Predictive Sample Reuse Method with Applications," *Journal of the American Statistical Association*, **70**, 320-328.
- Habbema, J. D. F., Hermans, J. and van den Broek, K. (1974) "A Stepwise Discriminant Analysis Program Using Density Estimation." In *Compstat 1974* ed. G. Bruckmann, pp. 101-110. Physica Verlag: Vienna
- Hall, P. (1984). "Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators," *Journal of Multivariate Analysis*, **14**, 1-16.

- Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Application*. Academic Press: New York.
- Hall, P. and Marron, J. S., (1987). "Extent to which Least-Squares Cross-Validation Minimizes Integrated Square Error in Nonparametric Density Estimation," *Theory of Probability and Related Fields*, **74**, 567-581.
- Härdle, W., Hall, P., and Marron, J. S. (1988) "How Far are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" with discussion, *Journal of the American Statistical Association*, **83**, 86-95.
- Härdle, W., and Marron, J. S. (1985) "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *Annals of Statistics*, **13**, 1465-1481.
- Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Linear Models*. Chapman and Hall: New York.
- Herrmann, E., Gasser, T., Kneip, A. (1992) "Choice of Bandwidth for Kernel Regression when Residuals are Correlated," *Biometrika*, **79**, 783- 795.
- Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statistics*, **19**, 293-325.
- Jones, M. C., Marron, J. S. and Park, B. U. (1991) "A Simple Root n Bandwidth Selector," *Annals of Statistics*, **19**, 1919-1932.
- Li, K.-C. (1985). "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-validation," *Annals of Statistics*, **13** 958-1377.
- Li, K.-C.(1986) "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing," *Annals of Statistics*, **14** 1101-1112.
- Li, K.-C. (1987) "Asymptotic Optimality for C_p , C_L , Cross-Validation and Generalized Cross-Validation: Discrete index set," *Annals of Statistics*, **15** 958-975.
- Mallows, C. (1973) " Some Comments on C_p ." *Technometrics*, **15**, 661-675.
- Milnor, J. (1963) *Morse Theory*, Princeton University Press, Princeton, New Jersey.
- Nadaraya, E. A. (1964) "On Estimating Regression," *Theory of Probability and its Applications*, **9**, 141-142.

- Priestley, M. B. and Chao, M. T. (1972) "Non-parametric Function Fitting". *Journal of the Royal Statistical Society B*, **34**, 385-392.
- Schuster, E. F. and Gregory, G. G. (1981) "On the Nonconsistency of Maximum Likelihood Nonparametric Density Estimators," in *13th Annual Symposium on the Interface of Computer Science and Statistics* ed. W. F. Eddy, pp. 295-298. New York: Springer-Verlag.
- Stone, C.J. (1977) "Consistent Nonparametric Regression," *Annals of Statistics*, **5**, 595-645.
- Stone, M. (1974) "Cross-Validatory Choice and Assessment of Statistical predictions," *Journal of the Royal Statistical Society Series B*, **39**, 44-47.
- Wahba, G. (1977) "Practical Approximate Solutions to Linear Operator Equations when the Data are Noisy," *SIAM Journal of Numerical Analysis*, **14**, 651-667.
- Wahba, G. (1983) "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistics Society, Series B*, **45**, 133-150.
- Wahba, G. (1990) *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, CBMS-NSF Regional Conference Series in Applied mathematics 59.
- Wahba, G. and Wold, S. (1975) "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics*, **4**, 1-47.
- Watson, G.S. (1964) "Smooth Regression Analysis," *Sankhya, Series A*, **26**, 359-372.
- Whittle, P. (1960) "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables", *Theory of Probability and Its Applications*, **5**, 302-305.
- Wong, W. H. (1983a) "A Note on the Modified Likelihood for Density estimation," *Journal of the American Statistical Association*, **78**, 461-463.
- Wong, W. H. (1983b) "On the Consistency of Cross-Validation in Kernel Nonparametric Regression," *Annals of Statistics* **78**, 1136-1141.

Gasser Muller kernel



Nadaraya-Watson kernel

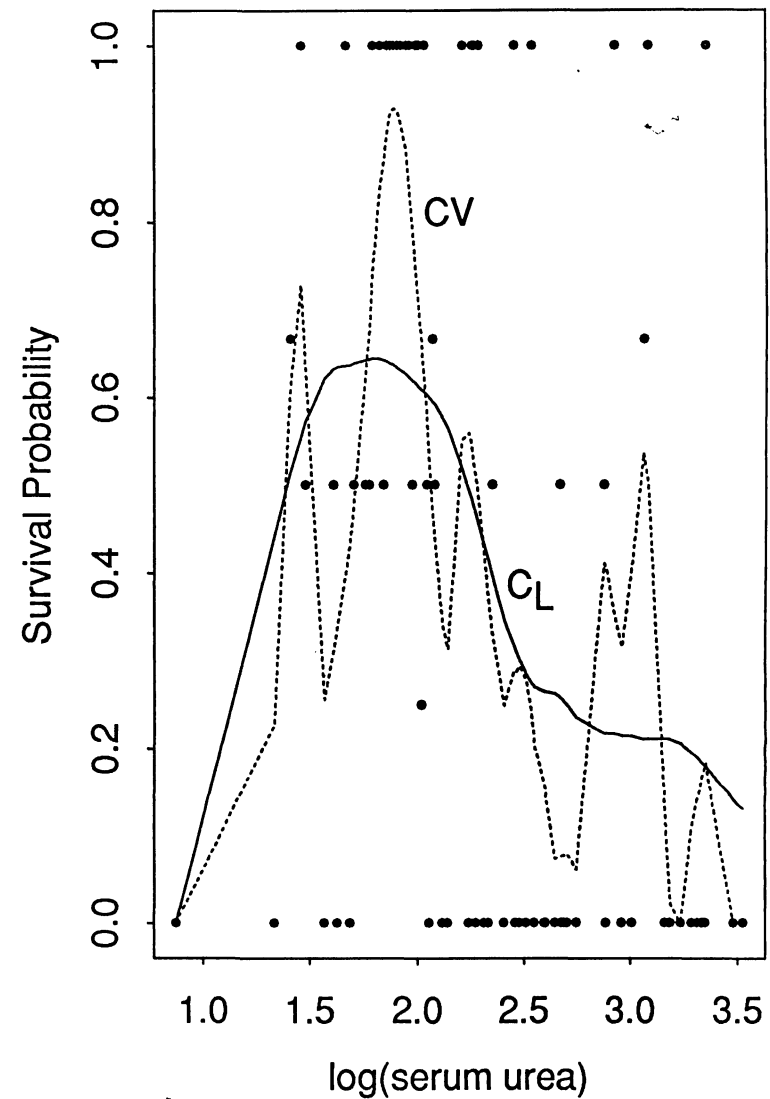


FIGURE 1: Kernel regression smooths of survival probability of periparturient recumbent cows as a function of log(serum urea), with bandwidth selected by CV and C_L . The dots denote the raw data.

Gasser Muller kernel

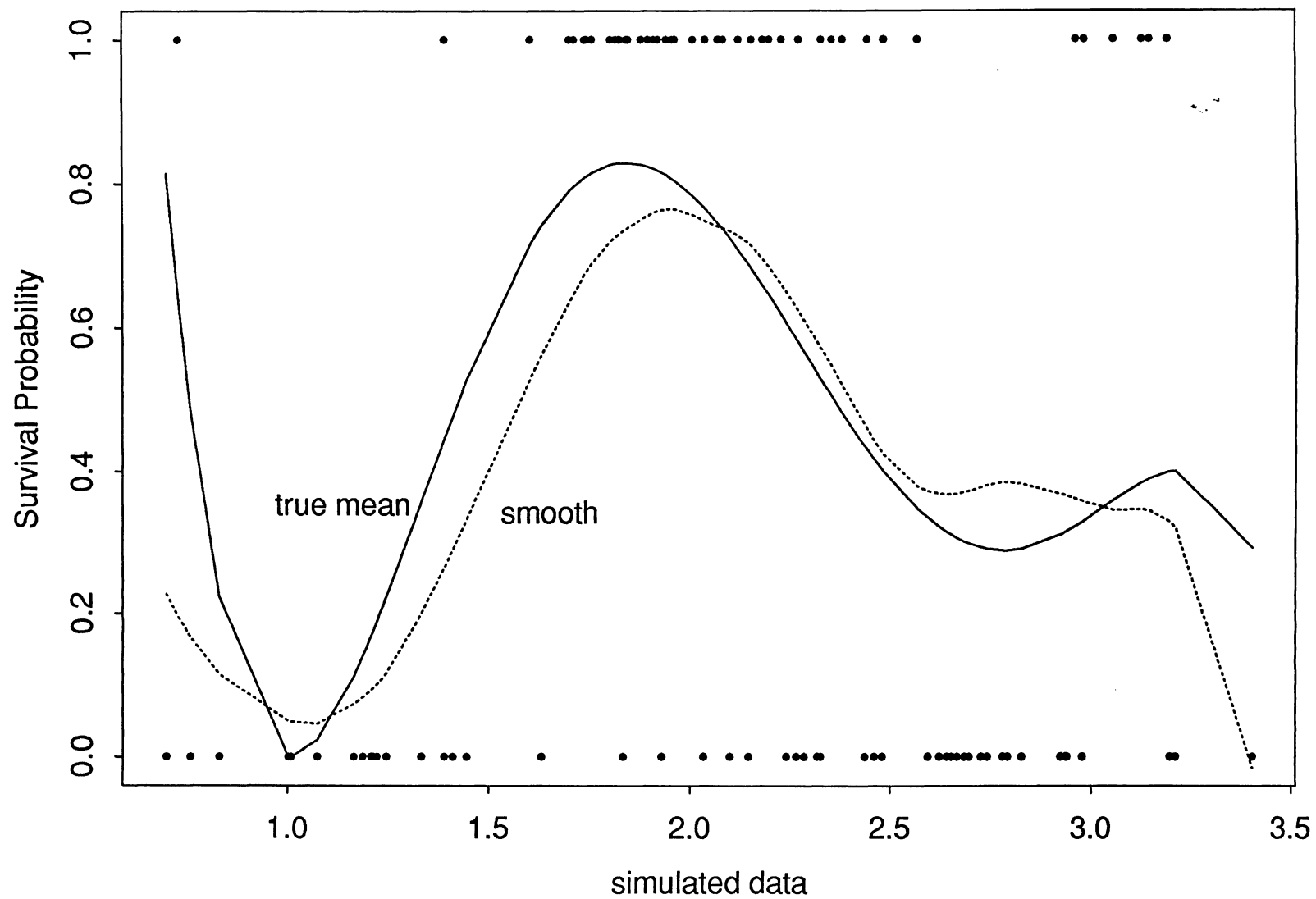


FIGURE 2: True mean function (solid line), and a kernel fit (dashed line) to data (dots) simulated to resemble the periparturient cow data.

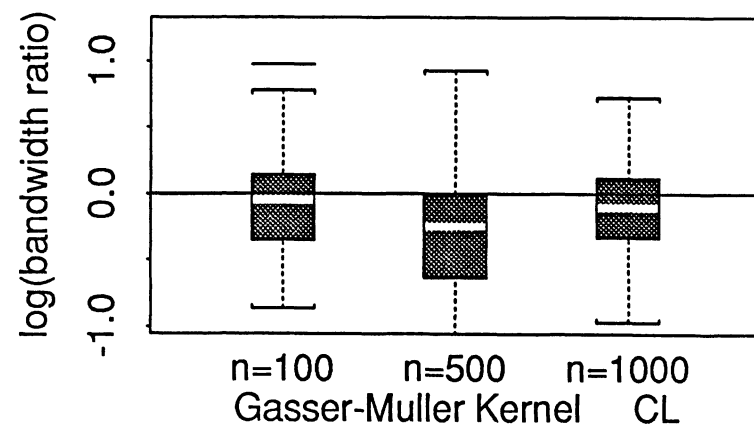
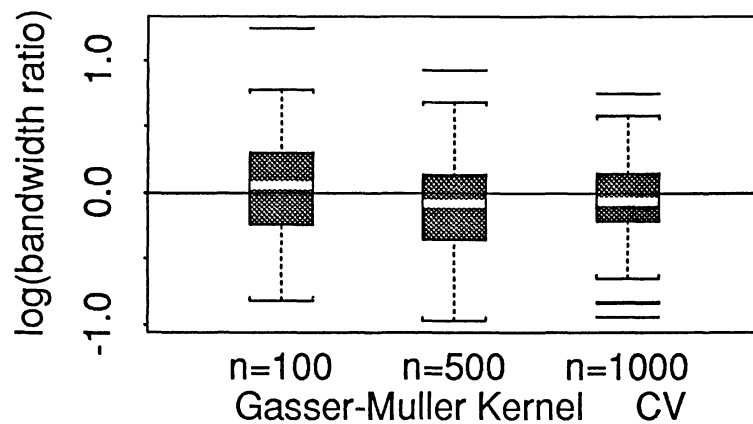
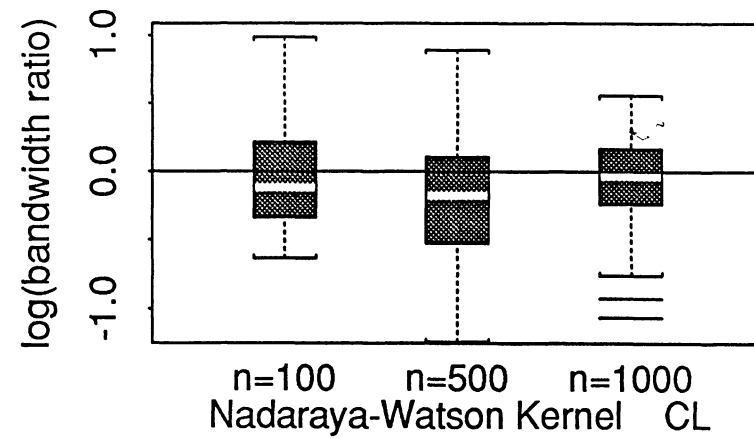
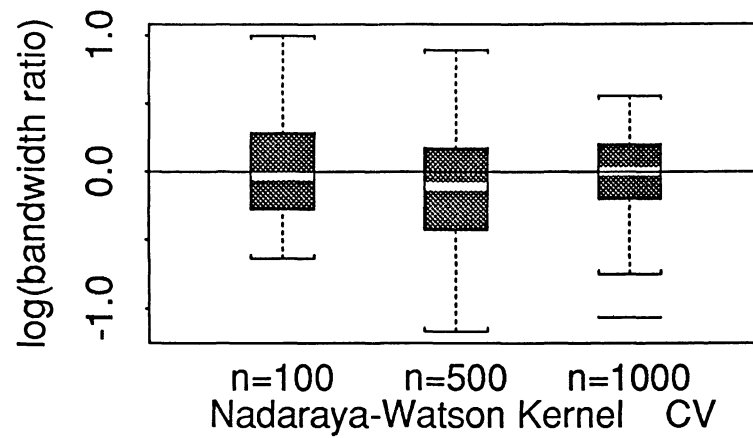


FIGURE 3: Comparison of selected to optimal bandwidth at various sample sizes for Nadaraya-Watson and Gasser-Muller kernels and bandwidth chosen by CV and CL. The y-axis is the logarithm of the ratio of selected to optimal bandwidth.

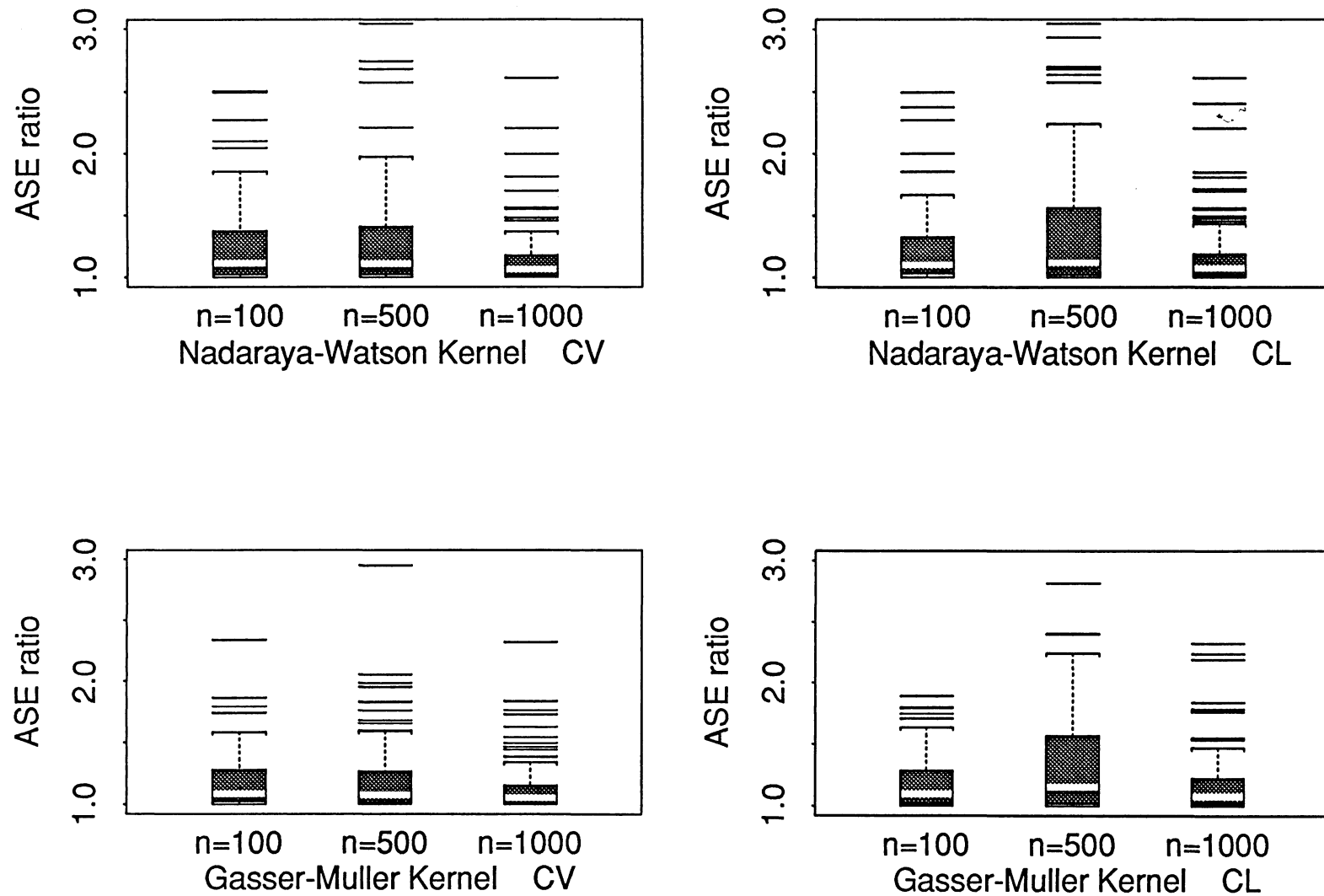


FIGURE 4: Comparison of ASE for selected and optimal bandwidth at various sample sizes for Nadaraya-Watson and Gasser-Muller kernels and bandwidth chosen by CV and CL. The y-axis is the ratio of ASE at the selected bandwidth to ASE at the optimal bandwidth.